SMU Data Science Review

Volume 3 | Number 2 Article 5

2020

Compressed DNA Representation for Efficient AMR Classification

John Partee
SMU, jpartee@smu.edu

Robert Hazell
Southern Methodist University, rhazell@mail.smu.edu

Anjli Solsi
asolsi@smu.edu

John Santerre
Sothern Methodist University, john.santerre@smu.edu

DNA Compression for Efficient Antibiotic Classification

Robert Hazell, John Partee, Anjli Solsi, and John Santerre

Master of Science in Data Science, Southern Methodist University,
Dallas TX 75275 USA
{rhazell, jpartee, asolsi, jsanterre}@smu.edu

Abstract. In this paper, we explore a representation methodology for the compression of DNA isolates. Using lossless string compression via tokenization of frequently repeated segments of DNA, we reduce the length of the isolates to be counted as k-mers for classification. With this new representation, we apply a previously established feature sampling method to dramatically reduce the feature space. In understanding the genetic diversity, we also look at conserving biological function across these spaces. Using a random forest model we were able to predict the resistance or susceptibility of bacteria with 85-90% accuracy, with a 30-50% reduction in overall isolate length, and an 80-90% reduction in the feature space over baseline. Significant contributions were built upon previous analysis of similar data.

1 Introduction

Antibiotics are drugs used to slow down or destroy bacteria. They have played a pivotal role in medical advances since the advent of Penicillin in 1928. Antibiotics have enabled the treatment of cancer, chronic disease, and surgery recovery through their ability to combat common infections [20]. The drawback is that as bacteria are exposed to antibiotics, resistance to the treatment can be developed for a variety of reasons.

The Centers for Disease Control and Prevention (CDC) has deemed antibiotic resistance a serious global threat. Over 2.8 million people in the United States become ill every year due to antibiotic-resistant infections, with roughly 35,000 people dying from those infections [7]. The CDC is responsible for creating awareness of the severity of the problem through stewardship programs such as the Antibiotic Resistant (AR) Solutions Initiative, investing in healthcare and community infrastructure to detect, respond, contain, and prevent resistant infections.

K-Mer analysis has been proven to be an effective way to identify antibiotic resistance in bacteria using machine learning, but has logistical problems that are difficult to overcome without considerable computing infrastructure. The theoretical feature space grows exponentially as the size of k increases. To combat this, we applied lossless string compression to DNA isolates, and utilized traditional k-mer analysis to the resulting string. The result was a dramatic

reduction in the number of unique k-mers across our samples, which brought with it a substantial reduction in k-mer counting times, model fitting times, and memory requirements, with only a modest drop in accuracy. Additionally, we observed that as the k-size increased, the time and memory requirements grew nearly linearly, instead of the typical exponential growth.

In this paper, we start by discussing the development of antimicrobial resistance and the concept of DNA compression utilized in the analysis. Next, various approaches that have been performed are reviewed in terms of scope, analysis techniques, and corresponding results. With this framework, we move onto our analysis, starting with an overview of the data, gathering the data set, and additional background. The analysis and results sections describe our compression method and models. Ethical considerations surrounding antimicrobial resistance are then presented, followed by the conclusions reached through this research and possible future work.

2 Background

2.1 Antimicrobial Resistance (AMR)

Antimicrobial resistance typically occurs when microorganisms develop a resistance after exposure to antimicrobial drugs. An example of a bacteria that has developed resistance is Neisseria Gonorrhea, which is now resistant to nearly all the antibiotics used for treatment. The identification of resistant genes in the DNA sequence of an organism, which are responsible for the genotype, is critical in understanding AMR [21]. A phenotype describes the observable characteristics of an organism that are related to its genotype, which is an organism's genetic composition. With this information, classification can be made as to whether antimicrobial bacteria are resistant or susceptible to antibiotics.

There are three methods for a bacterium to develop antimicrobial resistance. Bacterium can have a natural resistance to a drug. Second, resistant genes can be transferred among different species of bacteria through horizontal gene transfer (HGT). The third method is through the spontaneous occurrence of a mutation [20]. Mutations involve the presence of drugs removing sensitive genes and leaving resistant bacteria which is then passed on through natural selection [14]. Microorganisms are constantly evolving to find new defenses to survive the effects of drugs, called resistance mechanisms. Some of the resistance mechanisms include target alteration, impermeability, enzymatic modification or destruction, and efflux [18].

2.2 Neisseria Gonorrhea

For our research, we analyzed a collection of Neisseria Gonorrhea DNA isolates. Gonorrhea is an infection caused by a sexually transmitted bacterium. It is characterized by symptoms of discharge and inflammation of the urethra, cervix, pharynx, or rectum [20]. The prevalence is estimated at approximately 1.14

million new infections each year in the United States. The responsible bacteria, Neisseria Gonorrhea, has been assigned an urgent threat by the CDC, the highest level possible. The threat estimate for drug-resistant Neisseria Gonorrhea was 246,000 infections in 2013 and 550,000 infections in 2019. There was a reported 124 % increase in infections caused by Neisseria Gonorrhea in 2019 [7]. This infection causing bacteria has developed a resistance to all but one class of antibiotics, with ceftriaxone as the last recommended treatment.

2.3 K-mer Analysis

K-mers are commonly used in the field of bioinformatics, specifically sequence analysis, and are unique subsequences of length k. In this analysis, the k-mers referred to are comprised of nucleotides (A, T, G, and C), the building blocks of a DNA sequence. To better illustrate the concept, the following sample DNA sequence will be broken down into possible k-mers.

$$ACGTGACACT$$

Table 1. Example of possible k-mers from above sample DNA sequence.

k-value	k-mers
1	A, C, G, T, G, A, C, A, C, T
2	AC, CG, GT, TG, GA, AC, CA, AC, CT
3	ACG, CGT, GTG, TGA, GAC, ACA, CAC, ACT
4	ACGT, CGTG, GTGA, TGAC, GACA, ACAC, CACT
5	ACGTG, CGTGA, GTGAC, TGACA, GACAC, ACACT
6	ACGTGA, CGTGAC, GTGACA, TGACAC, GACACT
7	ACGTGAC, CGTGACA, GTGACAC, TGACACT

Once the DNA is broken down into K-mers, two major approaches can be used to create features for classification. Either the frequency of a K-mer in the sample can be counted, or the presence of a K-mer in a sample can be noted in a boolean fashion, with no concern given to recounts. For our research, we opted for the frequency-driven approach.

The k-mer approach was used by Lingle and Santerre [11] and is used in our analysis to find genetic features that are relevant in predicting whether a bacteria is susceptible or resistant in a random forest model. The choice of k is dependent upon balancing the effects. A lower k-value decreases the memory required for a sequence, while a larger k-value increases the storage and computation time since the feature space grows exponentially. However a larger k tends to produce more accurate models since it retains more of the sequence structure. For example, a k value of 7 in Table 1 yields k-mers more reflective of the original sequence.

3 Prior Analysis Approaches

We first consider some machine learning techniques used for AMR classification. To summarize from Lingle and Santerre [14], Davis et. al [8] explore bacteria genomes containing AMR data collected at PATRIC and model the antibiotic resistibility of several antimicrobials, including Acinetobacter baumannii, Staphylococcus aureus, Streptococcus pneumoniae, and Mycobacterium tuberculosis. This is accomplished through counting k-mers within each contig using the KMC counting software, then constructing a matrix configuration of the k-mers for each genome. The k-mers represent row and column values taking on binary 0/1 values representing the absence or presence, respectively, of each k-mer within a particular genome. Davis et. al settle on using 31-mers; determining the relevant k-mer markers for antibiotic resistance is performed using AdaBoost with 10 iterations of boosting. Davis et. al attain accuracies between 87-99% for Staphylococcus aureus and Streptococcus pneumoniae, while accuracies range from 71-88% for detecting AMR in Mycobacterium tuberculosis.

Lingle and Santerre [14] use a similar k-mer analysis and expand upon Davis et. al [8] by exploring other machine learning algorithms for AMR, including SVMs, Naïve Bayes, and random forests. Focus is on detecting AMR exclusively in Neisseria Gonorrhea from PATRIC. Unlike Davis et. al, analysis is restricted to k=5 through k=10 k-mers due to computational constraints. Lingle and Santerre's matrix structure also differs from Davis et. al in that the rows represent each isolate (rather than each k-mer) while column values represent the frequency with which a k-mer is present within an isolate (rather than a binary 0/1 representing the absence or presence of a k-mer). While computation increases significantly for 10-mers, ROC-AUC values are maximized with it, regardless of the model used. SVMs slightly outperform Random Forests (ROC-AUC of 0.94 vs 0.92) while Naïve Bayes is the worst performing overall, yielding a maximum AUC-ROC of 0.84.

Arango-Argoty et. al [2] use two deep learning models, DeepARG-SS and DeepARG-LS to determine antibiotic resistance genes (or ARGs) with data integrated from three genomic databases. The SS model intakes shorter DNA read sequences while the LS model is used for full gene sequences. These models were designed to improve upon current ARG identification by accounting for the similarity distribution of sequences in a way that reduces the false negative rate of AMR identification (e.g., incorrectly predicting a bacterium to be non-resistant to antibiotics). Arango-Argoty et. al achieve high precision (> 0.97) and recall (> 0.90) when testing their model on full length sequences. However, the SS model features high recall but low precision, implying a high false positive rate (precision$= 0.27$) for bacterium that are resistant to multiple drugs, like Mycobacterium tuberculosis.

Nguyen et. al [17] examine AMR in Salmonella genomes, considering how best to predict minimum inhibitory concentration (MIC) for 15 antibiotics. They take short-read sequence data from various strains and assemble 10-mers using the KMC program into a matrix with the k-mers, antibiotics, and MICs as features for each genome, where the rows took on k-mer values and the MIC. Similar

to Davis et.al [8] the matrix values take on 0/1 values denoting the absence or presence of a k-mer in the genome. XGBoost for regression is used to predict MIC for each antibiotic and determine the most important features for MIC prediction. Using 5,278 genomes all 15 antibiotics had average accuracies at or above 90%. However their models necessitated specialized hardware requirements of over 1.5 TB of ram and 22 cores of data, beyond the normal capacities of desktop computers.

Drouin et. al [9] consider Set Covering Machines (SCM) as an alternative to tree-based and SVM models, using 36 datasets with five bacterium (Acinetobacter baumanii, Mycobacterium tuberculosis, Pseudomonas aeruginosa, Staphylococcus aureus, and Streptococcus pneumoniae). The SCM intakes k-mers and a set of boolean-values rules that determine the presence or absence of a k-mer, finding the smallest set of rules that minimizes the probability of making a false prediction using a modification of the set covering algorithm. Drouin et. al examine the sensitivity-specificity tradeoff, achieving around 80% specificity for 33 of the 35 datasets and 80% specificity for 25 of the 36 datasets; the average number of rules derived was 2.5.

Eyre et. al [10] examine 681 Neisseria gonorrhea isolates from England, the USA, and Canada using whole genome sequencing and multivariate linear regression with interaction terms to predict MICs. The predictors comprised of several genetic determinants, including alleles. Five antimicrobials were tested (cefixme, penicillin, azithromycin, ciprofloxacin, and tetracycline); exact matches between predicted and observed MICs range between 44%-53% between the five antimicrobials. MIC values within one standard deviation of dilution doubling range between 91%-96% amongst the five antimicrobials. Overall pseudo-R^2 values ranged between 0.8-0.85, with ciprofloxacin featuring a 0.96 pseudo-R^2.

We next consider work in the compression or encoding of DNA and dimension reduction techniques. Cao et. al [6] develop a model termed an *expert model* (XM) and a DNA sequence compressor that encodes DNA base symbols based on calculating probabilities obtained from previous symbols in a sequence with a probability distribution. The "experts" are simply types of probability distributions for describing the occurrence of a DNA letter or sequence, such as Bayesian-based or Markov-based distribution. The XM model is able to compress the bits per symbol better than some compression methods (BioC and GenC) but less so compared to other methods like GeMNL and HUMHPRTB.

Al-Okaily et. al [1] investigate DNA compression of five different genomes through modification of Huffman encoding that allows for an unbalanced Huffman tree (UHT) that ensures three of the DNA bases to encoded using 2 bits and the fourth to be encoded using 3 bits, which improves on the current SHT encoding method. Compression ratios between 20%-27% are derived from UHT and UHTL encoding. MUHTL exhibits an approximate 3 percentage point reduction in file size compared to two popular file compression systems, gzip and bzip2, although MUHTL failed to show improvement with one of the genomes (Chr22) over bzip2 compression.

Though not related to AMR per se, Inza et. al [12] examine six filter methods for discrete data and one wrapper method (sequential forward selection – SFS) for continuous data for dimension reduction of DNA microarray data, comparing the LOOCV metric across four machine learning algorithms KNN, Naïve Bayes, Decision Trees, CN2 – a modification of Decision Trees) on two datasets for diagnosis of leukemia and colon cancer. All models show improvements in accuracy compared with that of no feature selection, selecting between two to four genes for the Colon dataset. Wrapper methods produced greater accuracies at the expense of higher computation times.

Benoit et. al [4] develop LEON, a lossless sequence compression algorithm that uses a probabilistic de Bruijn graph containing k-mers for each node, rather than a reference set of sequences, incorporating a Bloom filter for faster hash. LEON is tested on genomic data from E. coli, C. elegans, and a human genome and benchmarked against other compression tools like GZIP. LEON was able to compress the first two by a factor of 7 relative to the GZIP method, while LEON reduces the human genome file size from 733 GB to 47 GB, a reduction of 686 GB.

In the work surveyed, machine learning techniques for AMR are applied without consideration of compressing k-mer space. In our work we investigate various compression techniques and examine how much predictive accuracy is lost using such techniques.

4 Data

The data set used in this analysis originates from the Pathosystems Resource Integration Center, an information system at the University of Chicago funded by various Institutes, dedicated to providing integrated data and analysis tools to support the research community. The directory contains multiple folders with data for all public genomes. Through the FTP site, the .fna files pertaining to the Neisseria Gonorrhea bacterium can be downloaded. Within the Neisseria Gonorrhea antimicrobial resistance are subsets for azithromycin susceptibility and resistant isolates. Susceptibility indicates that the isolates can be treated with antibiotics, while resistance indicates the Gonorrhea are resistant to the treatment. There are 214 files for azithromycin resistant isolates and 183 files for azithromycin susceptible isolates used in this analysis.

Deoxyribonucleic acid (DNA) sequencing is a technique used to determine the order of nucleotides, commonly known as base pairs, which compose the foundation of a genome. The DNA sequence can be viewed as the blueprint of an organism, with the individual genes determining certain instructions. This view can assist in identifying certain genes that cause infections or disease. The four base pairs that bond the DNA strands together are Adenine (A), Cytosine (C), Guanine (G), and Thymine (T) [11].

The data files contain short reads, or fragments of DNA sequences of various lengths. These short reads are more commonly known as contigs. A contig is a DNA sequence that is assembled from overlapping shorter sequences to form

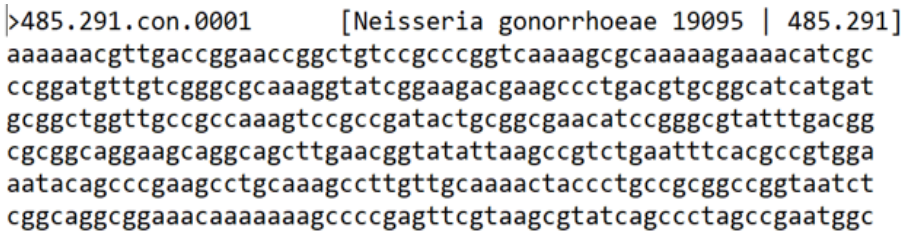

Fig. 1. An example of a Neisseria Gonorrhea isolate with the header information and part of the first contig.

one large contiguous sequence [5]. The lengths of the base pairs of each isolate vary and can range to over 2 million base pairs. The filenames indicate the strand number and genome id; however, those are not used for the purpose of the analysis.

5 DNA Compression Approach

DNA sequence information can require multiple gigabytes of storage. Data compression is a means of representing or storing data in such a way that reduces the storage size and/or length compared to the data's original encoding [1]. Compression methods can be classified as either lossy or lossless. With lossy encoding methods, the compressed version of data may eliminate some information considered non-essential. Lossless compression methods preserve all of the original data such that if the data were uncompressed, the exact information content from the original data representation would be present.

Besides lossy versus lossless compression, there are two modes for compressing information, DNA sequences in particular: horizontal mode and vertical mode [3]. Horizontal mode is more frequently used, and it entails compressing a DNA sequence using exclusively the information from that sequence alone – namely substrings of the sequence.

For our analysis, we performed compression via 'tokenization'. A 10% subsample of the DNA corpus was taken, and commonly repeated sequences of a preset token length within this subsample were iteratively counted and replaced by a single character, until the DNA had been reduced by a preset factor. Sequences of length 2-6 were tested, with compression factors of 30, 40, and 50 percent tried. Larger token lengths were tried, but were found to not be feasible, as the counts get more sparse as token lengths increase. Functionally, we are performing K-mer counting on a small random sample of the overall data, and replacing more common (thus, less useful) k-mers with a symbol.

As an example, to compress the sequence below with a token of length 4, to a factor of 50 percent, the following steps would be performed:

$$ACGACGTGCGACACGACGTGCGACTGTAC$$

Counting reveals that the most common subsequence of length 4 is 'CGAC'. This would be replaced with a symbol. For ease here, we will use the number 1. The resultant string would be:

$$A1GTG1A1GTG1TGTAC$$

whose length is reduced by 41%. For a compression factor of 40% (whose results are shown in the analysis and results sections), we would stop here. As this is below our preset stopping criteria, we'll do another replacement. After counting after the last replacement, the most common substring is 'A1GT'. In practice, it is uncommon to replace a sequence containing a token like this. Once replacing this substring with the number 2, we are left with the sequence:

$$2G12G1TGTAC$$

This is now reduced by 62%, so we would halt, and apply these two replacements across the rest of the data samples. Again, this is only a depiction of how the compression works. In practice, we typically overshot the compression target by 0.5-1% at most, and never replaced a series that had previously replaced tokens present.

6 Analysis

Our approach centers on implementing the compression algorithm outlined in the previous section to make a more compact representation of the isolates. The isolates in the data are further divided into k-mers of length k. The matrix representation of the data shows rows corresponding to the bacteria isolates, with columns representing unique k-mers found throughout all of the combined samples. The data themselves are the frequency of the given k-mer in each isolate.

As a random forest model has been shown to be effective on this data [14], we will use this as our benchmark model. Using parallelized k-mer counting and aggregation, and sparse matrices, we were able to replicate the accuracy of Lingle and Santerre's study [14], with quicker training and counting times. A chart of these processing times (counting and model training) for a 40% compression factor are below shown in figure 1. Further results for 30% and 50% compression are available in Appendix A.

For these baseline trials, we processed each isolate as a concatenated string, with all contigs considered together. We tested token lengths from 2 to 5, each with k-mer lengths of 5 to 15. To compress the isolates, we created a key from the set of the characters present in the isolate, which is the Cartesian product of that set, repeated by the token length. From there, we counted the occurrences of all of the candidate keys, and replaced the highest counted key with a token if it was more than one percent of the overall document. After each replacement, the tokens are recounted and replaced until a replacement produces a reduction of less than one percent.

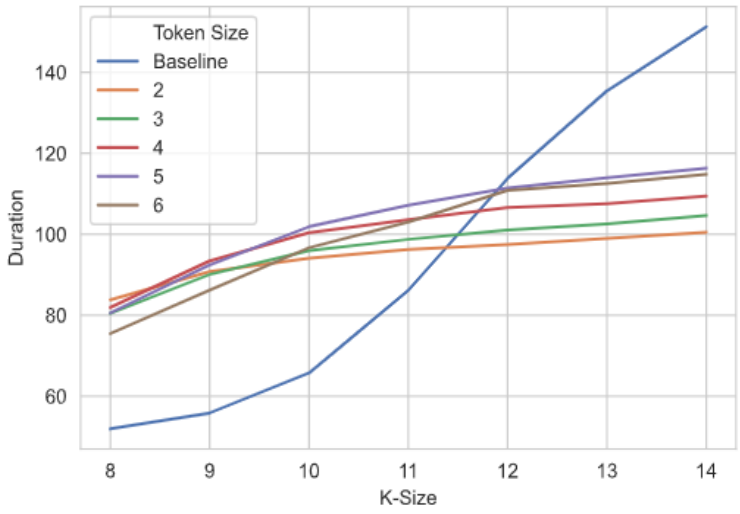

Fig. 2. Modeling times based on k-mer size and token length.

Early trials were carried out with tokenization replacing every set of bases for the token length, but the k-mer space size was too large for training times to be reasonable. K-mer space tends to be both sparse and concentrated, and this concentration was eliminated with tokenization. Our approach of only replacing series that were disproportionately represented reduced that problem of tokenization. As such, sequences present in the DNA that were less informative presumably are represented less in the resultant model.

Post tokenization and counting, we dropped all of the non-unique columns from the matrix, functionally collapsing identical columns into one, following with Santerre's Palaverous sampling method[19], and additionally removing any columns which were all of the same count. The efficiency gained here is shown below. The feature space grows similarly to the training time shown above; nearly linearly once the isolates are compressed. Figure 3 demonstrates this growth. Figure 4 shows the number of features that the model was actually trained on, post Palaverous sampling. The end result is a dramatic reduction in the feature space, which means smaller models, and faster training times. The figures below show the output of sampling post 40% compression, results at 30% and 50% are available in the appendix.

7 Results

As discussed above, the end result of representing data this way is that the feature space is substantially smaller, while reducing the initial isolate length substantially, pre-K-mer counting. This results in reduced memory requirements for pre-processing, and smaller models. As we can see, no compression is still

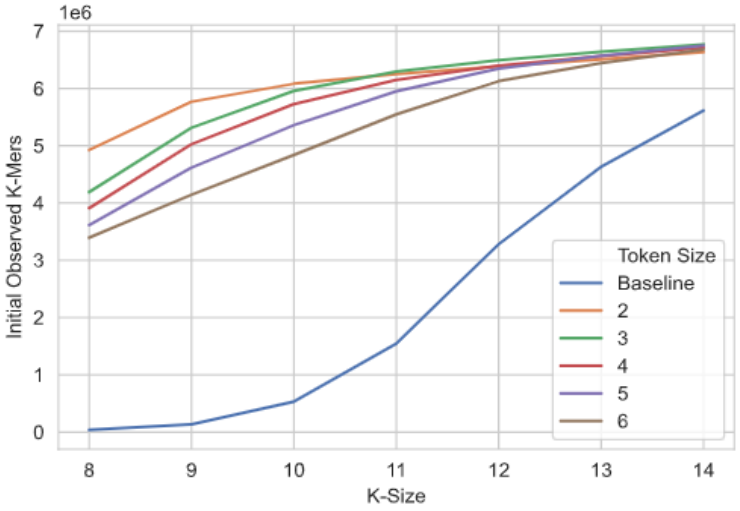

Fig. 3. Initial Unique K-Mers by K-size.

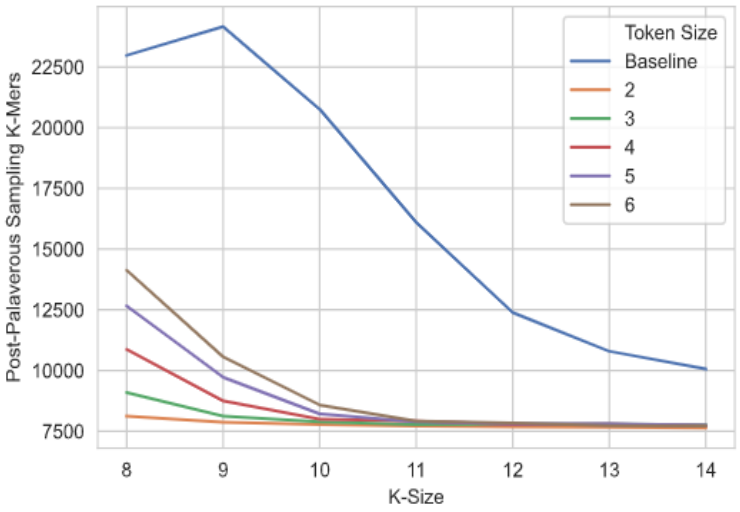

Fig. 4. Unique K-Mers post-Palaverous sampling.

more accurate over roughly K=11, but the reduction in training times shown above can outweigh that slight accuracy increase when speed is required.

The trial results are shown below in figure 5. The results presented here are using a 40% compression target, as above. Results at 30% and 50% are available in appendix A.

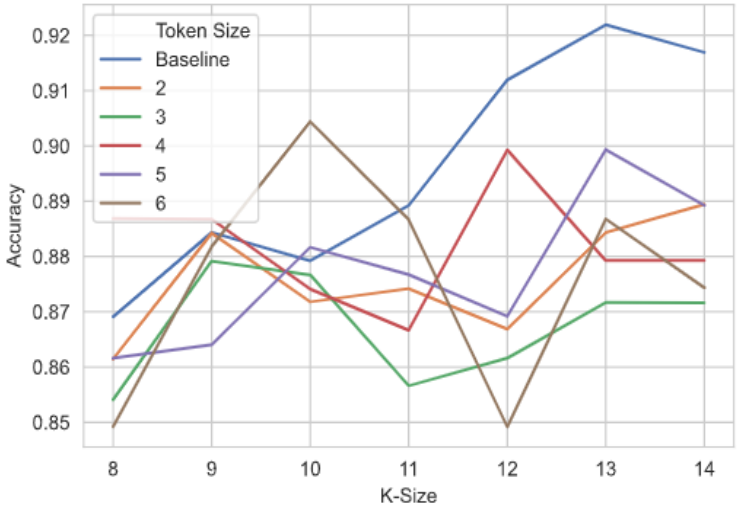

Fig. 5. Trial results of base random forest model displaying classification accuracy based on k-mer size and token length.

The initial problem with this approach was that more unique tokens meant more unique k-mers, and a larger feature space. This problem has been more or less eliminated with paleverous sampling. Once pre-processed, the model training time and memory requirements are dramatically reduced from baseline.

We believe that this increase in speed with only a modest reduction in accuracy is due to what the K-mers now represent. As an example, a K-Mer of length five, with token sizes of three, could represent up to 15 individual bases. This approach functionally creates variable length K-mer sizes, allowing more information to be embedded in some features than is traditionally available.

8 Conclusions

We found that the K-mer approach for DNA classification is incredibly effective. However, there is room for improvement in processing and memory efficiency. The traditional approach required significant memory, and processing power. We were able to replicate the accuracy of previous studies done on large computers with esoteric tools with a laptop and a set of tools that we have published on

PyPi. These efficiency increases aside, we were able to shorten the length of the isolates via lossless string compression, and using palaverous sampling we were able to further reduce the memory and compute power required to accomplish this analysis. The cumulative effect of these efficiencies is that we are able to perform what was once difficult computations on commodity hardware with little effort, enabling quicker analysis, at lower cost, opening the door to k-mer analysis being done in the field.

9 Future Work

This method focused on implementing a tokenization algorithm to make a more compact representation of k-mers. The output of the compression was used in modeling to identify gene regions attributed to antibiotic resistance. The compression algorithm created in this analysis could be rewritten in a manner that allows for faster run time. Other compression methods could be tested and compared to the results found here, such as Huffman coding, which is commonly used in lossless data compression. Other natural language processing techniques could be implemented in normalization and pre-processing of the data to improve accuracy. Different classifiers, such as hierarchical bayes, could be applied in the analysis. While the approach utilized in this analysis succeeded in effectiveness of DNA classification, the processing and accuracy can be improved upon. Additionally, the tools used in this research will be released in the near future under the PyPi package name, "OKmers", to provide the computational research community with more high speed parallel DNA tools.

10 Ethics

The ethics surrounding antimicrobial resistance vary across industries. Antibiotics are limited resources, and a main issue is the overuse and incorrect use of antibiotics. According to the CDC, as much as 50% of the time, antibiotics are wrongly prescribed for viral infections and other pathogens, such as colds or flus, or people are prescribed an incorrect dose [7]. As more antibiotics are prescribed and used, the opportunity for bacteria to develop resistance increases quickly. Certain infections, tonsillitis, gonorrhea, pneumonia, that were previously treatable with antibiotics are slowly becoming untreatable. Devastating effects could result from the abuse of antibiotics, such as an inability to utilize them for serious infections, surgeries, cancer treatment, transplants, and chronic diseases [20]. To combat the misuse of antibiotics the CDC has organized stewardship programs such as Detect and Protect against Antibiotic Resistance, and the government formed the National Strategy for Combating Antibiotic-Resistance Bacteria (CARB) to develop best practices and guidelines to assist in dissemination as well as working to implement data analysis to drive diagnostic testing.

Excessive and incorrect use of antibiotics is also present in the agriculture and farming industries, raising concerns over animal welfare and farming practices. Antibiotics are used frequently in food-producing animals, and resistant bacteria

can then be contracted by humans through animal consumption. An estimated 80% of antibiotics in the United States are used in animals to promote growth, increase feed efficiency, and prevent infection [20]. Resistant germs present in the guts of animals can contaminate meat and other products. Animal waste can also carry bacteria, and when used in fertilizer, other produce and organisms can be contaminated through contact with the soil and runoff water. The pesticides used in agriculture for crops can also contaminate irrigation water and further affect other water systems. Veterinary Feed Directives have been developed to combat this issue. The Food and Drug Administration (FDA) has approved antibiotic use in food-producing animals for the following reasons: treatment of disease, control of disease in a population when some animals are ill, and prevention of disease for at-risk animals [13].

The UK was the first nation to identify issues with livestock production reliance on antibiotics and created recommendations for therapeutic administration. This response was met with political opposition from farming and pharmaceutical industries, leading to difficulties in rollout and implementation. However, Sweden was the first country to ban the use of growth promoting antibiotics in all livestock [13].

The pursuit for drug development and discovery of new antibiotics has diminished in the past 30 years. This lack of research and development has led to a reliance on the same drugs for decades, leading to bacterial evolution and development of resistance. The general timeline of development and research to review new drugs takes over 6 years. It is a complex, costly process that has led pharmaceutical companies to focus on short-course therapies, drugs taken every day that combat chronic issues.

Another cause of increase in AMR is due to healthcare practices and sanitation uses. Faulty infection control protocols in hospitals and clinics make such places susceptible to the spread of antimicrobial resistance. A lack of rapid lab testing to determine diagnosis leads to diminished healthcare. Data collection on patients with infections and treatments needs to be shared globally to influence testing and policy decisions. Antibacterial household products used for hygiene and cleaning have been noted as a cause of AMR and may limit development of immunity in children [16].

A different ethical approach considers economic and sociological factors. One area of concern is the restriction of individual liberties due to the emergence of drug resistance infections [15]. This can be related to the Covid-19 pandemic and varying opinions and opposition to the different lockdown and quarantine rules made with a focus on public health. Globally, low-income countries lack the resources and awareness, while antibiotics are unregulated, available over the counter in other countries [16].

References

1. Al-Okaily, A., Almarri, B., Yami, S.A., Huang, C.H.: Toward a better compression for dna sequences using huffman encoding. Journal of Computational Biology **24**(4), 280–288 (April 2017), doi: 10.1089/cmb.2016.0151

2. Arango-Argoty, G., Garner, E., Pruden, A., Heath, L.S., Vikesland, P., Zhang, L.: Deeparg: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome **6**(23) (February 2018), doi: 10.1186/s40168-018-0401-z
3. Bakr, N.S., Sharawi, A.A.: Dna lossless compression algorithms: Review. American Journal of Bioinformatics Research **3**(3), 72–81 (2013), doi:10.5923/j.bioinformatics.20130303.04
4. Benoit, G., Lemaitre, C., Lavenier, D., Drezen, E., Dayris, T., Uricaru, R., Rizk, G.: Reference-free compression of high throughput sequencing data with a probabilistic de bruijn graph. BMC Bioinformatics **16**(288), 91–10 (September 2015), doi: 10.1186/s12859-015-0709-7
5. Binder, M., Hirokawa, N., Windhorst, U. (eds.): Contig, p. 871. Springer-Verlag Berlin Heidelberg, Berlin, Germany, 1 edn. (2009)
6. Cao, M.D., Dix, T.I., Allison, L., Mears, C. (eds.): A Simple Statistical Algorithm for Biological Sequence Compression. IEEE Computer Society (March 2007), doi: 10.1109/DCC.2007.7
7. CDC: Antibiotic resistance threats in the united states. Tech. rep., U.S. Department of Helath and Human Services, CDC, Atlanta, GA (2019)
8. Davis, James J, e.a.: Antimicrobial resistance prediction in patric and rast. Scientific Reports **6**(27930) (June 2016)
9. Drouin, A., Raymond, F., Letarte, G., Marchand, M., Corbeil, J., Laviolette, F.: Large scale modeling of antimicrobial resistance with interpretable classifiers (December 2016), arXiv:1612.01030v1
10. Eyre, D.W., Silva, Dilrini De, e.a.: Wgs to predict antibiotic mics for neisseria gonorrhoeae. Journal of Antimicrobial Chemotherapy **72**(7), 1937–1947 (July 2017), doi: 10.1093/jac/dkx067
11. Heather, J.M., Chain, B.: The sequence of sequencers: The history of sequencing dna. Genomics **107**(1), 1–8 (January 2016), doi: 10.1016/j.ygeno.2015.11.003
12. Inza, I., Larrañaga, P., Blanco, R., Cerrolaza, A.J.: Filter versus wrapper gene selection approaches in dna microarray domains. Artificial Intelligence in Medicine **31**(2), 91–103 (May 2004), doi: 10.1016/j.artmed.2004.01.007
13. Kahn, L.H.: Antimicrobial resistance: a one health perspective. Transactions of The Royal Society of Tropical Medicine and Hygiene **111**(6), 255–260 (June 2017), doi: 10.1093/trstmh/trx050
14. Lingle, J.I., Santerre, J.: Using machine learning for antimicrobial resistant dna identification. SMU Data Science Review **2**(2), Article 12 (2019)
15. Littmann, J., Buyx, A., Cars, O.: Antibiotic resistance: An ethical challenge. International Journal of Antimicrobial Agents **46**(4), 359–361 (October 2015), doi: 10.1016/j.ijantimicag.2015.06.010
16. Littmann, J., Viens, A.M.: The ethical significance of antimicrobial resistance. Public Health Ethics **8**(3), 209–224 (November 2015), doi: 10.1093/phe/phv025
17. Nguyen, M., Long, S.W., McDermott, P.F., Olsen, R.J., Olson, R., Stevens, R.L., Tyson, G.H., Zhao, S., Davis, J.J.: Using machine learning to predict antimicrobial mics and associated genomic features for nontyphoidal salmonella. Journal of Clinical Microbiology **57**(2) (January 2019), doi: 10.1101/380782
18. Poole, K.: Mechanisms of bacterial biocide and antibiotic resistance. Journal of Applied Microbiology **92**(s1), 55S–64S (May 2002)
19. Santerre, J.W., Davis, J.J., Xia, F., Stevens, R.: Machine Learning for Antimicrobial Resistance. Ph.D. thesis, University of Chicago, Chicago, IL (July 2016), arXiv preprint arXiv:1607.01224

20. Ventola, C.L.: The antibiotic resistance crisis: part 1: causes and threats. P & T : a peer-reviewed journal for formulary management **40**(4), 277–283 (April 2015)
21. Zankari, Ea, e.a.: Identification of acquired antimicrobial resistance genes. Journal of Antimicrobial Chemotherapy **67**(11), 2640–2644 (November 2012), doi: 10.1093/jac/dks261

Appendix A

1. Results at 30% Compression

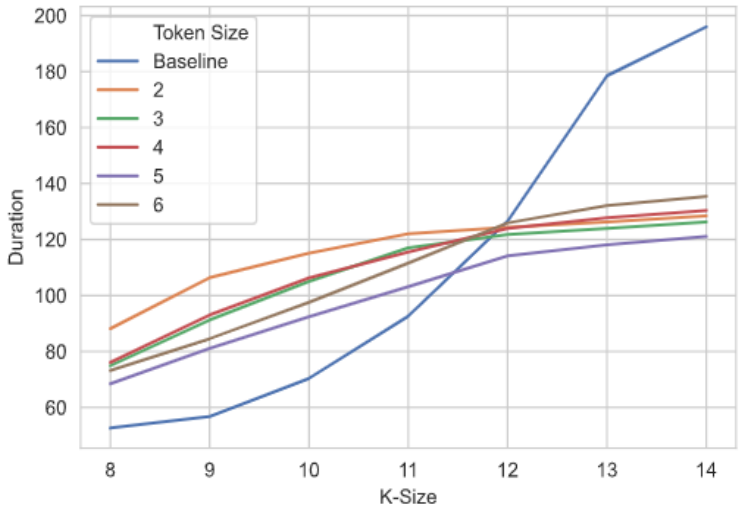

Fig. 6. Modeling times based on k-mer size and token length.

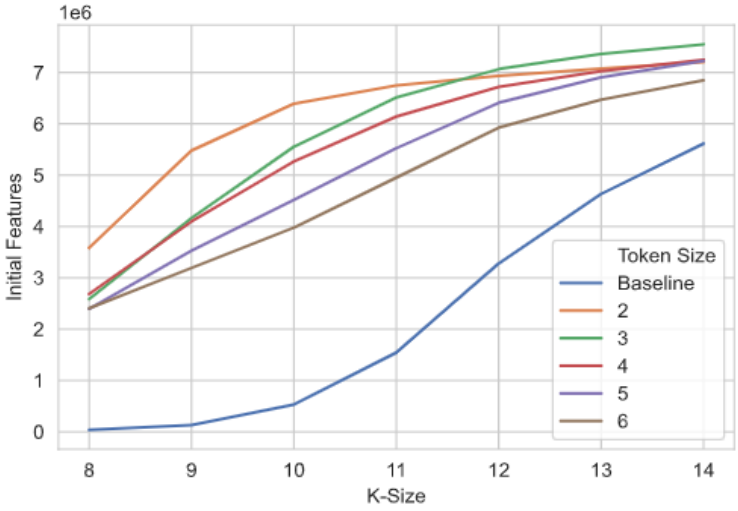

Fig. 7. Initial Unique K-Mers by K-size.

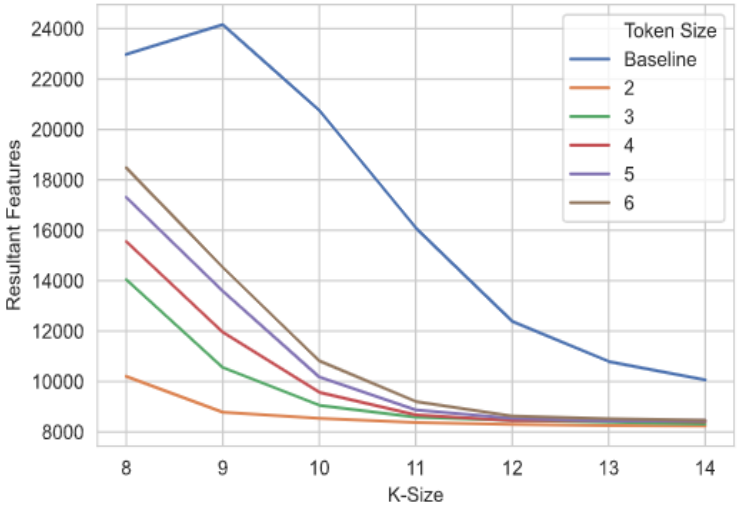

Fig. 8. Unique K-Mers post-Palaverous sampling.

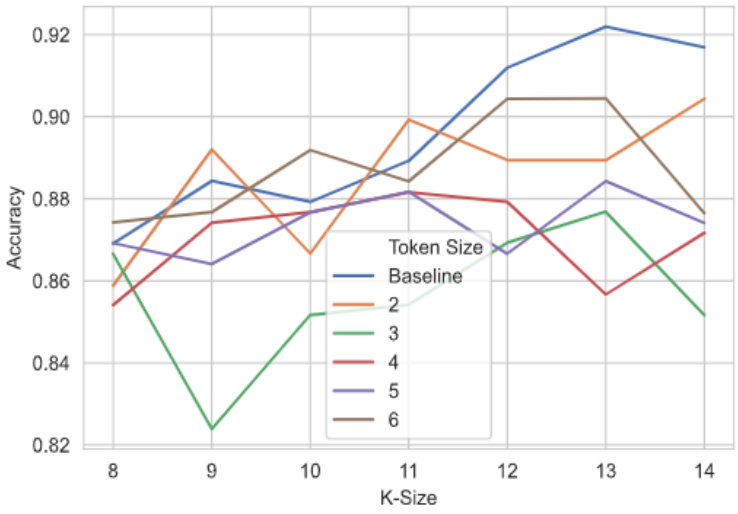

Fig. 9. Trial results of base random forest model displaying classification accuracy based on k-mer size and token length.

2. Results at 50% Compression

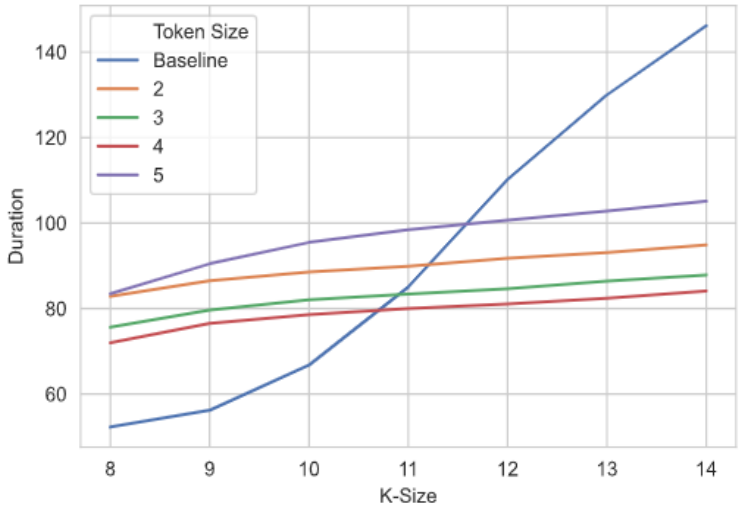

Fig. 10. Modeling times based on k-mer size and token length.

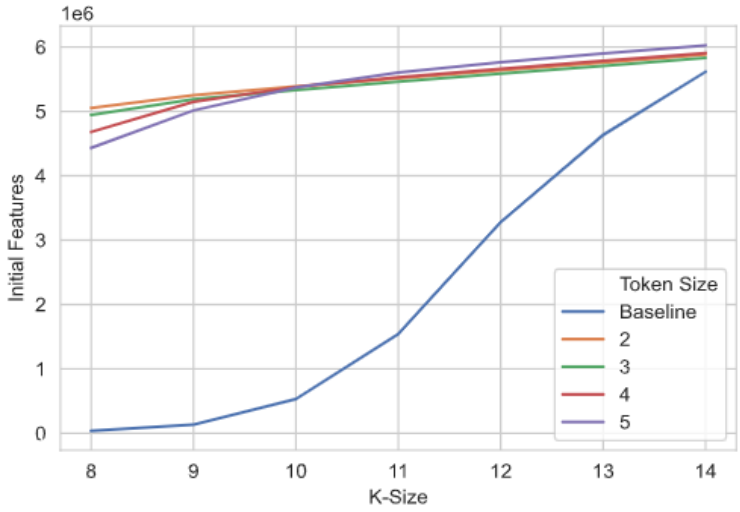

Fig. 11. Initial Unique K-Mers by K-size.

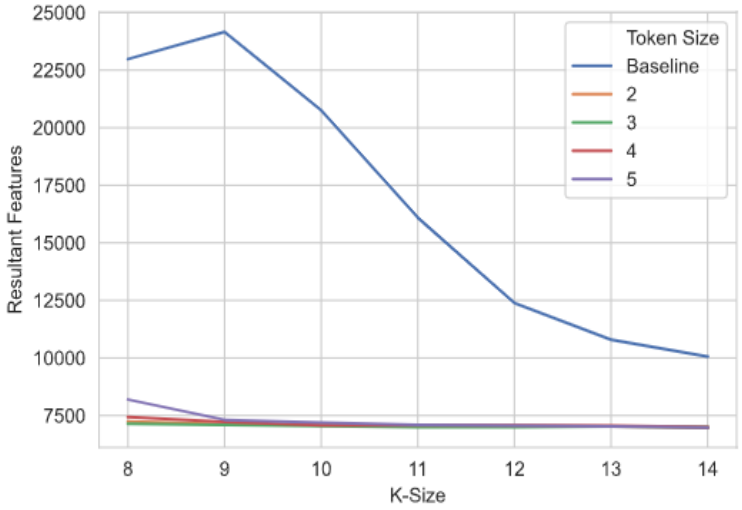

Fig. 12. Unique K-Mers post-Palaverous sampling.

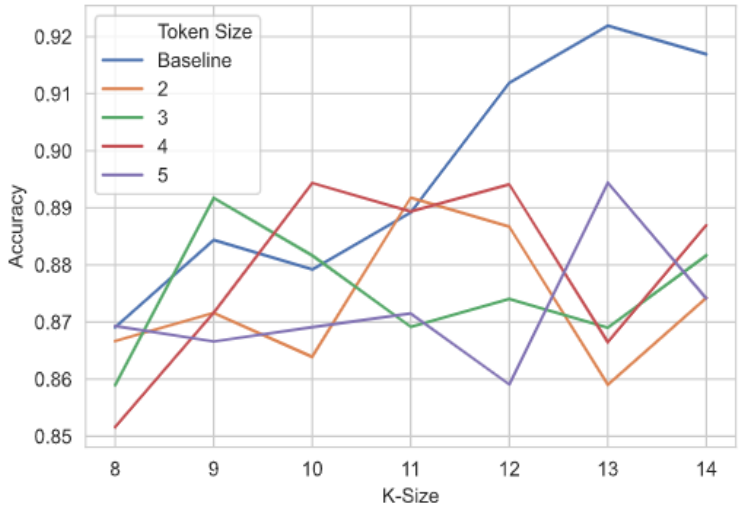

Fig. 13. Trial results of base random forest model displaying classification accuracy based on k-mer size and token length.